What is claimed is:

1. A method of managing data storage in a data processing apparatus, the data processing apparatus including an information repository comprising a physical data storage medium and data structures for storing index information for locating data in the data storage medium, the method comprising the steps of:

analyzing the contents of the set of files to identify components of the file contents which have duplicates within different files within the set;

deleting duplicate components from the information repository while retaining at least one copy of each component, and generating index data for the retained copies which reflects the respective logical positions within the information repository corresponding to the positions of the retained copies and their deleted duplicates, and generating index data for remainder components which correspond to the remainder portions of a file after separation of duplicated components which remainder component index data reflects the logical positions of the remainder components within the information repository; and

storing the generated index data.

2. A method according to claim 1, wherein the analysis of file contents comprises the steps of:

separating file contents into a set of information components comprising sub-sections of a file's contents, in accordance with predefined separation criteria; and

analyzing the contents of said information components to identify duplicates.

3. A method according to claim 2, wherein the step of separating a file's contents into information components is initiated in response to a step of saving the file, and the steps of analyzing the contents to identify duplicates and then deleting duplicates are performed by a background process independently of user-controlled operations.

4. A method according to claim 2, wherein said step of separating file contents comprises identifying a file type, selecting predefined separation criteria for the identified file type, and separating file contents in accordance with the selected separation criteria.

5. A method according to claim 1, including the step of identifying information components corresponding to sub-sections of an identified component of a file's contents, which sub-sections have duplicates within different files within the set, and performing in relation to said sub-section components said steps of deleting duplicates and generating and storing index data for retained single copies of duplicated sub-section components and generating and storing separate index data for remainder sub-section components.

6. A method according to claim 5, wherein said steps of deleting duplicates and generating separate index data is performed subject to a defined minimum component size.

7. A method according to claim 1, wherein the generated index data comprises:

a set of file descriptions which each include an ordered list of identifiers of components corresponding to the contents of the respective file and information defining a path within a directory structure corresponding to the logical location of the file within the directory structure; and

a set of unique component identifiers to be stored in association with respective components.

8. A method according to claim 7, wherein the index data is implemented using markup tags, with each unique component identifier comprising a unique tag pair identifying and delimiting the respective component within the information repository and said ordered list of component identifiers within each file description comprising a list of markup tags.

9. A method according to claim 7, wherein the index data additionally comprises:

an indication of the locations within the information repository of members of said set of unique component identifiers.

10. A data processing apparatus comprising:

an information repository for storing a set of files and for storing index information for locating files within the information repository; and

controller components for controlling the operation of the data processing apparatus to perform the following method steps:

analyzing the contents of a set of files stored in the information repository to identify components of the file contents which have duplicates within different files within the set;

deleting duplicate components from the information repository while retaining at least one copy of each component, and generating index data for the retained copies which reflects the respective logical positions within the information repository corresponding to the positions of the retained copies and their deleted duplicates, and generating index data for remainder components which correspond to the remainder portions of a file after separation of duplicated components which remainder component index data reflects the logical positions of the remainder components within the information repository; and

storing the generated index data.

11. A data processing apparatus according to claim 10, wherein the controller component for generating index data is adapted to generate:

a set of file descriptions, which each include an ordered list of identifiers of information components

corresponding to the contents of the respective file and information defining a path within a directory structure corresponding to the logical location of the file within the directory structure; and

a set of unique component identifiers to be stored in association with respective components;

wherein the apparatus further comprises a component for analysing the index data for all components of the set of files to identify and generate a representation of a directory structure.

12. A data processing apparatus according to claim 10, including a publish/subscribe engine connected for communication between application programs and said controller components for analyzing contents, deleting duplicates and generating indexes, wherein the publish/subscribe engine enables the application programs to register as publishers and as subscribers for information and is adapted to compare information components created by a first application program with other application programs' subscriptions, and then to notify identified subscriber applications when a created information component matches an application program's subscriptions.

13. A data processing apparatus according to claim 10, including one or more search agents for performing search and retrieval operations from the information repository in response to requests from one or more application programs.

14. A computer program product comprising program code recorded on a computer-readable recording medium, the program code including instructions for controlling the operation of a data processing apparatus, when executed thereon, to perform a method for managing storage of a set of files within an information repository, the information repository comprising a physical data storage medium and data structures for storing index information for locating files in the physical data storage medium, wherein the program code comprises:

means for analyzing the contents of the set of files to identify components of the file contents which have duplicates within different files within the set;

means for deleting duplicate components from the information repository while retaining at least one copy of each component, and for generating index data for the retained copies which reflects the respective logical positions within the information repository corresponding to the positions of the retained copies and their deleted duplicates, and for generating index data for remainder components which correspond to the remainder portions of a file after separation of duplicated components which remainder component index data reflects the logical positions of the remainder components within the information repository; and

means for storing the generated index data.